

ARMY RESEARCH LABORATORY



Arabic Natural Language Processing System Code Library

by Stephen C. Tratz

ARL-TN-0609

June 2014

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory

Adelphi, MD 20783-1197

ARL-TN-0609**June 2014**

Arabic Natural Language Processing System Code Library

Stephen C. Tratz

Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
<p>Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.</p> <p>PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.</p>					
1. REPORT DATE (DD-MM-YYYY) June 2014		2. REPORT TYPE Final		3. DATES COVERED (From - To) 10/2013–09/2014	
4. TITLE AND SUBTITLE Arabic Natural Language Processing System Code Library				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Stephen C. Tratz				5d. PROJECT NUMBER R.0010376.10	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U.S. Army Research Laboratory ATTN: RDRL-CII-T Adelphi, MD 20783-1197				8. PERFORMING ORGANIZATION REPORT NUMBER ARL-TN-0609	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT <p>This technical note provides a brief description of a Java library for Arabic natural language processing (NLP) containing code for training and applying the Arabic NLP system described in the paper "A Cross-Task Flexible Transition Model for Arabic Tokenization, Affix Detection, Affix Labeling, POS Tagging, and Dependency Parsing" by Stephen Tratz presented at the Statistical Parsing of Morphologically Rich Languages (SPMRL) workshop held in Seattle in conjunction with the Empirical Methods in Natural Language Processing (EMNLP) conference of October 2013. The system is capable of clitic separation, inflectional affix identification and labeling, part-of-speech tagging, and dependency parsing for Arabic. The code, which is extended from previously released graduate student code, also supports English part-of-speech tagging, dependency parsing, and semantic disambiguation tasks. In general, the code library is expected to be of most value to natural language processing researchers.</p>					
15. SUBJECT TERMS Arabic, natural language processing, NLP, Java, code					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES 14	19a. NAME OF RESPONSIBLE PERSON Stephen C. Tratz
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			19b. TELEPHONE NUMBER (Include area code) (301) 394-2305

Contents

1. Introduction	1
2. File Overview	1
3. Code Compilation	2
4. Training Instructions	2
5. Applying the System to New Examples	2
6. License	3
7. History	3
8. Important Note	4
9. Papers to Cite	4
Distribution List	7

INTENTIONALLY LEFT BLANK.

1. Introduction

This technical note provides a brief description of a Java library for Arabic (and also English) natural language processing (NLP), containing code for training and applying the Arabic NLP system described in Stephen Tratz’s 2013 paper “A Cross-Task Flexible Transition Model for Arabic Tokenization, Affix Detection, Affix Labeling, POS Tagging, and Dependency Parsing” in the 2013 Statistical Parsing of Morphologically Rich Languages (SPMRL) workshop (see section 9 for all relevant publications).

The system is capable of clitic separation, inflectional affix identification and labeling, part-of-speech tagging, and dependency parsing for Arabic. For more information on these topics and the overall system, the reader is referred to the SPMRL 2013 publication.

For historical reasons (section 7), the library also includes components for English part-of-speech tagging, dependency parsing, preposition sense disambiguation, noun compound interpretation, and possessives interpretation. Much of the code is relevant to general purpose NLP applications (e.g., the WordNet interface) and is not specific to either Arabic or English.

2. File Overview

The following is a short description of the top level files and directories.

- build/ – directory with Ant build scripts for running dependency converters, training the various subsystems, and running some experiments
- conf/ – directory with files related to Arabic constituent-to-dependency conversion as well as the English semantic annotation modules
- data/ – directory with a variety of data used by different submodules
- docs/ – directory with additional documentation
- examples/ – a directory with some example input and output files
- jar/ – a directory with a precompiled copy of this code
- lib/ – a directory for any required .jar files
- scripts/ – a directory with some scripts for applying trained models to new data
- src/ – the source code
- README – a README file giving an overview similar to this technical note

LICENSE – the license file

finegrainconverter-0.2.3.jar

– English dependency converter jar from
<http://sourceforge.net/projects/miacp/>

pennconverter_modified_Spring2012.jar

– English dependency converter jar from
<http://sourceforge.net/projects/miacp/>
(modified from the pennconverter
http://nlp.cs.lth.se/software/treebank_converter/)

3. Code Compilation

The code is typically compiled using Apache Ant, freely available software typically used for building software projects. To download, or for more information, see <http://ant.apache.org/>.

A variety of ant build XML files are contained in the build subdirectory. To create a jar file containing the compiled code, switch to INSTALLATION_DIR/build and type “ant -f build_common.xml makeJar”.

4. Training Instructions

For the Arabic system, see doc/AR_TRAINING_INSTRUCTIONS.

For the English system, see doc/EN_TRAINING_INSTRUCTIONS .

5. Applying the System to New Examples

Example scripts for applying a trained model to new data are provided under the scripts subdirectory. Short example texts are provided in the examples subdirectory.

6. License

This project constitutes a work of the United States Government and is not subject to domestic copyright protection under 17 USC Â§ 105.

The project includes and derives from code licensed under the terms of the Apache 2.0 License and is, therefore, licensed under the Apache 2.0 License. See the doc/LICENSE-2.0 file for more information. This file may also be found at <http://www.apache.org/licenses/LICENSE-2.0>.

A copy of WordNet 3.0 (<http://wordnet.princeton.edu/wordnet/>), a freely available lexical database for English, is included with this project. Its license is provided in the ‘doc/WORDNET_LICENSE’ file.

A modified version of the pennconverter, freely available constituent-to-dependency conversion software described by Richard Johansson and Pierre Nugues in their paper “Extended Constituent-to-dependency Conversion for English,” in the proceedings of the 2007 Nordic Conference on Computational Linguistics, is included with this project. Its license is provided in the ‘doc/PENNCONVERTER_LICENSE’ file.

7. History

This project is derived from a collection of code written as part of my graduate student/thesis work at USC/ISI that was publicly released under the Apache 2.0 license in late 2011 at <http://www.isi.edu/publications/licensed-sw/fanseparator/index.html>.

In mid-2012, a project derived from the original University of Southern California/Information Sciences Institute version of the software was released at <http://sourceforge.net/project/s/miacp/>. This newer version used a slightly different English dependency scheme and contained a variety of improvements. However, the PropBank-style SRL module was not maintained.

This current project was derived from the <http://sourceforge.net/projects/miacp/> release by NLP researchers at the U.S. Army Research Laboratory. A significant portion of the code has been replaced with newer code, deleted, or otherwise revised. The focus of this effort was to create the system described in the SPMRL 2013 paper mentioned in section 9. However, the English POS

tagger, parser, preposition sense disambiguator, noun compound relation disambiguator, and possessive relation disambiguator are still functional. (The English components may be slightly slower and slightly less accurate than those in the <http://sourceforge.net/projects/miacp/> release.)

8. Important Note

This release contains a variety of bug fixes and other generally small changes that have been made since the SPMRL 2013 paper publication and, as such, results will not be identical—in general, the results will be insignificantly better, but there may be some cases where performance is worse.

9. Papers to Cite

Authors wishing to cite to this work should refer to the following papers.

Arabic

Tratz, Stephen. 2013. *A Cross-Task Flexible Transition Model for Arabic Tokenization, Affix Detection, Affix Labeling, POS Tagging, and Dependency Parsing*. Fourth Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL).

English

(Note: These are for earlier versions of the English system/components and, thus, it may make more sense to cite one or more of these along with the 2013 SPMRL paper.)

Tratz, Stephen. 2011. Ph.D. Thesis. *Semantically-Enriched Parsing for Natural Language Understanding*. University of Southern California.

Dependency Parsing

Tratz, Stephen, and Eduard Hovy. 2011. *A Fast, Accurate, Non-projective, Semantically-Enriched Parser*. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

Preposition Sense Disambiguation

Tratz, Stephen, and Dirk Hovy. 2009. *Disambiguation of Preposition Sense Using Linguistically Motivated Features*. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Companion Volume: Student Research Workshop and Doctoral Consortium.

Hovy, Dirk, Stephen Tratz, and Eduard Hovy. 2010. *What's in a Preposition?: Dimensions of Sense Disambiguation for an Interesting Word Class*. Proceedings of the 23rd International Conference on Computational Linguistics (COLING): Posters.

Noun Compound Interpretation

Tratz, Stephen, and Eduard Hovy. 2010. *A Taxonomy, Dataset, and Classifier for Automatic Noun Compound Interpretation*. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL).

Possessives Interpretation

Tratz, Stephen, and Eduard Hovy. 2013. *Automatic Interpretation of the English Possessive*. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL).

INTENTIONALLY LEFT BLANK.

<u>NO. OF COPIES</u>	<u>ORGANIZATION</u>
1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
2 (PDF)	DIRECTOR US ARMY RESEARCH LAB RDRL CIO LL IMAL HRA MAIL & RECORDS MGMT
1 (PDF)	GOVT PRINTG OFC A MALHOTRA
1 (PDF)	DIRECTOR US ARMY RESEARCH LAB RDRL CII T S TRATZ

INTENTIONALLY LEFT BLANK.